ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Dynamic Self-Distillation Transformer is Used for Fault Detection of Charging Piles with Limited Characteristics

Yifei Li[1,2,*], *Yutong Zhao*[1,2], *Jing Shen*[1], *Meiying Yang*[1,2], *Tianle Li*[1,2] *and Bin Zhao*[1,2], *Yi Li*[2], *Xinzhi Lin*[2] *and Zhao Zhang*[2]

[1] State Grid Beijing Electric Power Company Electric Power Science Research Institute, BeiJing, 100031, China
[2] Beijing Dingcheng Hong'an Technology Development Co., Ltd, BeiJing, 100075, China

SUMMARY: *With the rapid increase in the number of electric vehicles, the safety and stability issues of electric vehicle charging equipment are becoming increasingly prominent. The failure of charging equipment not only affects the user experience but may also lead to economic losses and safety hazards. The development of efficient and accurate charging pile fault detection algorithms is of great significance. Based on this, this paper proposes a dynamic self-distillation transformer for fault detection of charging piles with limited features. The method first reconstructs the limited features through permutation and combination, effectively increasing the representation ability of the input features. Then, the transformer network is used for feature extraction of the new features. To reduce the complexity of model deployment, this paper designs a dynamic self-distillation method to compress the knowledge of deeper transformer networks into shallower networks, effectively enhancing the detection performance of the shallower network. Experiments conducted on real datasets have proven that this method achieves higher fault detection accuracy than existing advanced models. It also demonstrates that the dynamic knowledge distillation strategy can not only reduce computational power for deployment but also better learn the knowledge of deeper networks to improve the detection accuracy of the model.*

KEYWORDS: *charging pile fault detection, reconstructs the limited features dynamic self-distillation, transformer networ*

1 Introduction

Against the backdrop of the deepening global consensus on addressing climate change and promoting green transportation, the electric vehicle industry is showing an exponential growth trend [1, 2]. In 2023, global electric vehicle sales will exceed 14 million units, accounting for nearly 18% of total new vehicle sales, and will continue to grow rapidly in the next decade. By 2040, the global electric vehicle ownership will exceed 700 million units. As a key interface connecting the power grid and electric vehicles, the reliability, safety, and operational efficiency of the charger are directly related to user experience, power grid safety and stability, and high-quality industrial development. However, in complex and changing operating conditions, electric vehicle charging piles face various potential failure risks [3].

The main types of faults in electric vehicle chargers include power module failure, DC contactor adhesion or burning, charging connection device overheating or damage, abnormal

*liyifei@bj.sgcc.com.cn
https://doi.org/10.65102/is2026834
Published: 30 April, 2026

insulation performance, communication link interruption, and software logic abnormalities. Charging machine failures not only reduce users' charging experience, cause resource waste and increase operating costs, but also pose safety hazards [4], and high-frequency failures will weaken users' trust in electric vehicles and charging services. Therefore, building an intelligent and highly responsive electric vehicle charger fault diagnosis system has important practical significance, which can significantly improve the efficiency and service quality of the charging network operation, and is an important support for ensuring the sustainable and healthy development of the electric vehicle industry.

In recent years, domestic and foreign scholars' research on fault diagnosis technology for electric vehicle chargers has gradually developed from traditional diagnosis methods based on rules and thresholds in the early days to intelligent fault prediction and diagnosis systems that integrate cutting-edge technologies such as signal processing, machine learning, artificial intelligence, big data analysis, and the Internet of Things (IoT). For example, various improved heuristic algorithms have been used in related studies to optimize the support vector machine (SVM) used for charging fault detection, which significantly improves the accuracy of fault recognition compared to traditional SVM models. In addition, research has delved into the application effects of ensemble learning methods such as random forest (RF) and XGBoost in the evaluation of charging station operation status. Reference [5] proposes a new method for fault detection in charging stations based on deep learning. By analyzing the output voltage waveform of the rectifier, effective identification of fault states in power devices and control units is achieved, which improves the accuracy and real-time performance of the detection process. Reference [6] uses a one-dimensional convolutional neural network (1D-CNN) to mine the characteristics of charging voltage and current signals; Reference [7] introduces Long Short Term Memory (LSTM) networks to adapt to the temporal characteristics of the charging process. Reference [8] extracts key fault features through wavelet packet transform and optimizes the BP neural network using sparrow search algorithm, effectively improving the accuracy and speed of fault diagnosis, and can directly output fault classification results, achieving efficient diagnosis goals. Wavelet transform and S-transform are widely used for fault feature extraction of non-stationary charging signals due to their time-frequency analysis characteristics. For example, reference [9] combines the efficiency of affinity propagation (AP) clustering algorithm with the precise classification ability of hidden Markov model (HMM) to construct an APHMM hybrid diagnostic model, which not only improves diagnostic accuracy, but also applies to electronic device fault analysis scenarios that require high diagnostic accuracy. Bayesian network and other probability graph models have also shown great potential for application in the field of complex system fault diagnosis. For example, reference [10] proposes a hybrid network structure that combines Convolutional Neural Networks (CNN) and Long Short Term Memory Networks (LSTM) to improve model training performance.

The field of fault diagnosis for charging stations faces three core challenges: 1) Due to sensor deployment costs, data collection and transmission constraints, and commercial privacy limitations, the available operational data for charging stations is generally scarce [11], making it difficult for them to effectively learn deep fault modes; 2) Although deep learning models have superior performance, they generally suffer from problems such as complex structures, large parameter quantities, and high computational costs, making it difficult to directly deploy and achieve real-time online diagnosis; 3) There are significant differences in the data distribution of charging stations of different brands, models, and operating environments, which puts strict requirements on the generalization of the model.

In this regard, the academic community has conducted research on various data augmentation techniques: (1) constructing effective features based on prior knowledge and

signal processing methods, or using variational autoencoders (VAEs) and other generative models to construct simulation data to expand training samples; (2) Using model compression and knowledge distillation (KD) to transfer knowledge from complex teacher models to lightweight student models [12]. Self distillation (SD) is an improved paradigm for knowledge distillation, which can extract knowledge during the training process without pre training the teacher model. The dynamic distillation strategy can further improve the efficiency of knowledge transfer. In addition, the Transformer model, with its excellent global dependency modeling ability and efficient parallel computing characteristics, has shown great potential for application in the analysis of timing data and fault diagnosis of electromechanical equipment in recent years, providing a new technological path for complex timing pattern recognition [13].

This article proposes a fault detection model for charging stations in feature constrained scenarios based on Dynamic Self Distillation Transformer (DSD Transformer). Firstly, the limited original features of the charging station are reconstructed and upgraded through permutation and combination methods to enhance the expression ability of input features and alleviate the problem of feature scarcity; Secondly, relying on the efficient feature extraction capability of the Transformer network, the enhanced features are deeply mined and a dynamic self distillation mechanism is designed to transfer and compress the knowledge of the deep Transformer network to the shallow network, enabling the shallow model to maintain high-precision fault detection capability. Research has shown that the proposed model can simultaneously address the two key challenges of insufficient feature expression and lightweight deployment of complex models.

2 Charging pile data acquisition sensor system

The charging pile data acquisition sensor system mainly collects the following operational data [14, 15]: K1K2 access control signal (x1), electronic lock drive signal (x2), emergency stop voltage value (x3), access voltage value (x4), harmonic distortion voltage value (x5), and harmonic distortion current value (x6).

The sensor system consists of five core modules: power supply unit, metering unit, ADC sampling unit, MCU control unit, and data storage unit. Among them, the function of the power unit is to supply power to the entire system; The function of the measuring unit is to achieve real-time acquisition of the waveform of the output voltage and current of the charging pile; The function of the ADC sampling unit is to complete the analog-to-digital conversion and sampling processing of the low-voltage signal of the charging pile; The function of MCU control unit is to undertake data processing and logic control; The function of the data storage unit is to store the collected signals into a storage medium, providing data support for subsequent data analysis and fault diagnosis.

The principal block diagram of the charging pile data acquisition sensor system is illustrated in Figure 1.

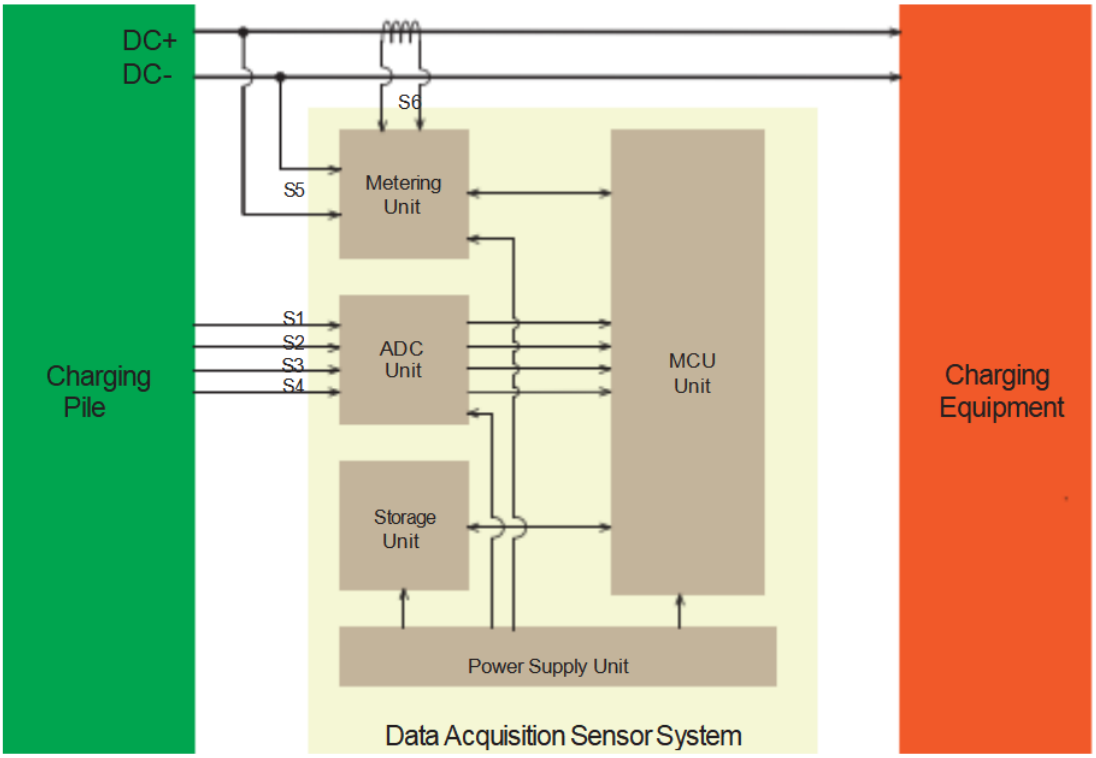

Figure 1: The principle block diagram of the charging pile data acquisition sensor system

The collection of signals S1 through S4 from the charging pile is primarily accomplished by the ADC sampling unit for front-end sampling, with the MCU control unit facilitating analog-to-digital conversion and data processing. Voltage signals are protected against EMC interference by passing through safety capacitors, self-resetting fuses (PTC), transient voltage suppressors (TVS), and gas discharge tubes (GDT) to prevent static electricity or induced voltages from damaging the sampling unit. After protection, the S1 to S4 signals are divided using high-precision, low-drift sampling resistors, filtered by RC filters, and then input into ADC channels ADC1 to ADC4 of the MCU control unit. To improve the accuracy of data collection and suppress electromagnetic interference under high voltage and high power conditions, the ADC sampling data is calculated using a sliding filtering algorithm to obtain the effective value. The ADC1 to ADC4 channels of the MCU control unit complete equivalent synchronous sampling in timer triggered scanning mode. The collected data is stored in a two-dimensional array format, and the sampling sequences of each channel contain ADC values and timestamp information, providing data support for the analysis of voltage timing relationships between nodes S1 to S4 [16].

The collection process of S5 and S6 is more complex and requires cooperation between the metering unit and the main control unit. Due to the high voltage, large current, and high power of the original signals, direct connection to the sampling unit is avoided to ensure personal safety. Isolation and voltage division methods are employed for collection. Specifically, the current signal of the charging pile is first isolated and processed with a current ratio by a high-precision (0.02% linearity), high-bandwidth (over 500KHz) zero-flux current transformer, with an isolation withstand voltage reaching the 4KV level. The voltage signal is divided using multiple series-connected high-precision, low-drift MELF resistors to increase the withstand voltage capacity, obtaining a low-voltage signal from the maximum 1KV DC voltage (DC+, DC-). The current ratio and voltage division signals are then input into the metering chip. The metering chip converts the analog signal to a digital signal at a

sampling frequency of 6.4kHz, and the digital signal is transmitted to the MCU control unit via the SPI interface, thereby obtaining the original waveform data of voltage and current. The original waveform data is transformed into frequency-domain signals through Fourier transform, represented as the superposition of voltage and current components at different frequencies. Taking the measurement of voltage THD as an example, the Fourier series representation formula is as follows [17]:

$$v(t) = v_1 + \sum_{n=2}^{\infty} \left[v_n \cos\left(nwt + {}_n \right) \right] \tag{1}$$

where *v1* is the DC voltage component of the charging pile, *v2 ~ vn* are the amplitudes of the corresponding higher harmonic voltages, w is the angular frequencies of the higher harmonic voltages, and *n* is the phases of the corresponding higher harmonic voltages.

The Fast Fourier Transform (FFT) algorithm efficiently calculates the frequency-domain signals corresponding to continuously sampled time-domain signals. By obtaining the DC voltage component and the amplitudes of different harmonic voltage components, the voltage THD can be calculated using the following:

$$THD = \sqrt{\sum_{n=2}^{50} v_n^2} \Big/ v_1 \tag{2}$$

The calculation method of THD for current is analogous to that for voltage.

3 Methodology

The issue of fault detection in charging piles lies in utilizing the limited current and voltage signals obtainable from the equipment to establish a complex correlation. By means of this correlation, we can determine whether a charging pile has malfunctioned. The dynamic knowledge self-distillation Transformer method proposed in this paper serves as the aforementioned correlation. Its specific structure is depicted in Figure 2 and consists of an attribute reconstruction module, a feature extraction network, and a dynamic knowledge distillation module. The following paragraphs will provide an in-depth explanation of this methodology [18].

3.1 Feature extraction network

In actual operation, EV chargers may collect limited data like current and voltage due to technical and cost limits, which may not fully mirror their operational status. To better present features, this paper reconstructs limited features via permutation and combination. As shown in Figure 2, assuming the input feature dimension is R1×n and the permutation and combination window is m, the reconstructed feature dimension becomes $R_m \times C_n^{\mathrm{m}}$, where:

$$C_n^m = \frac{n!}{m!(n-m)!} \tag{3}$$

After enhancing features through permutation and combination-based reconstruction, this paper proposes an adaptive feature fusion method rooted in a channel attention mechanism. By allocating learnable weights to the channels of reconstructed features, this approach

empowers the model to dynamically learn the relevance of each feature channel for fault detection, achieving weighted feature integration.

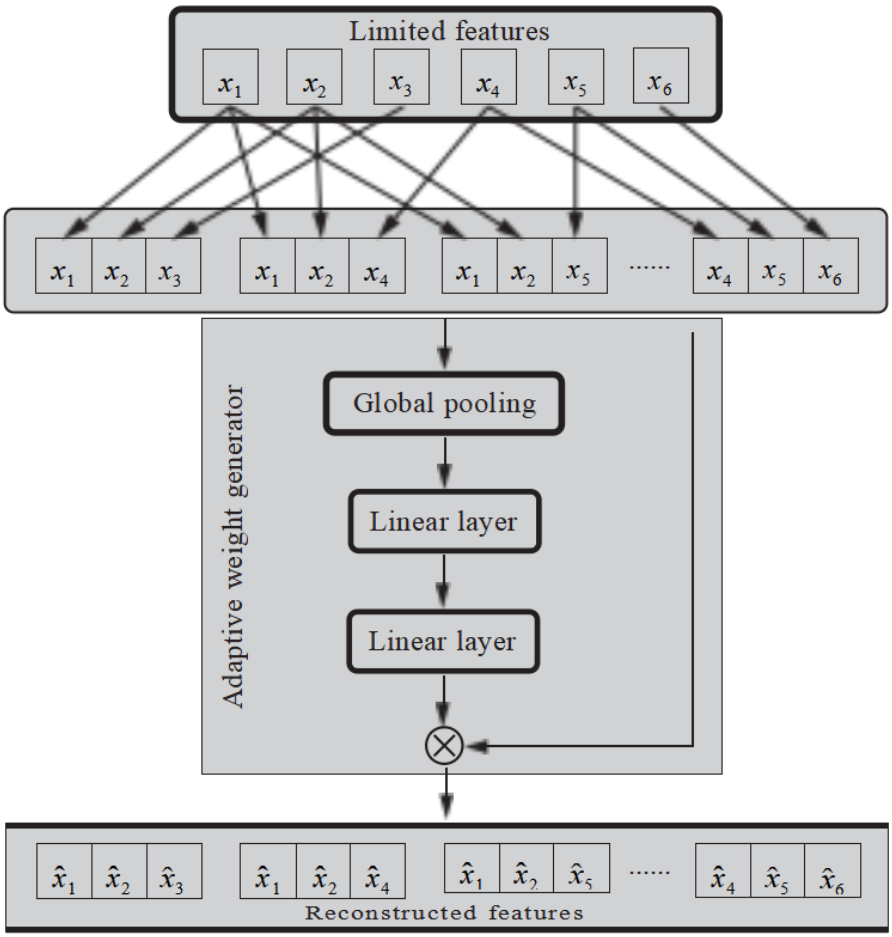

Figure 2: Adaptive feature reconstruction flowchart

Firstly, implement a mean pooling process on the re-constructed feature $X \in R_m \times C_n^{\mathrm{m}}$ to create a channel-wise statistical descriptor vector s:

$$s_c = \frac{1}{m}\sum_{i=1}^{m} X_{i,c}; c = 1,2,\cdots,C_n^{\mathrm{m}} \tag{4}$$

where s_c represents the global statistical features of the cth channel.

Then, feed the statistical vector s into a two-layer linear network to generate channel-wise weights via nonlinear transformation:

$$W = (s \cdot W_1 + b_1)W_2 + b_2 \tag{5}$$

where W_1 and W_2 represent the learnable parameter matrices of the linear layer, r is the compression ratio, (.) is the Silu function and (.) is the sigmoid function, which is used to normalize the weights to the interval of [0, 1].

Subsequently, the channel weights w are multiplied with the original reconstructed feature X on a per-channel basis to obtain the adaptively fused features [19]:

$$\hat{X}_{i,c} = X_{i,c} \cdot W_c \tag{6}$$

By incorporating a channel weighted mechanism, the model can adaptively learn feature importance during end-to-end training. This enables it to focus more effectively on key features and improves the feature extraction capability of subsequent Transformer modules.

After obtaining the reconstructed input features, multiple layers of Transformer blocks are employed to extract the hidden information of the reconstructed features. Each Transformer block consists of a multi-head self-attention network, Root Mean Square (RMS) normalization, a feedforward network, and residual shortcuts, as illustrated in Figure 3.

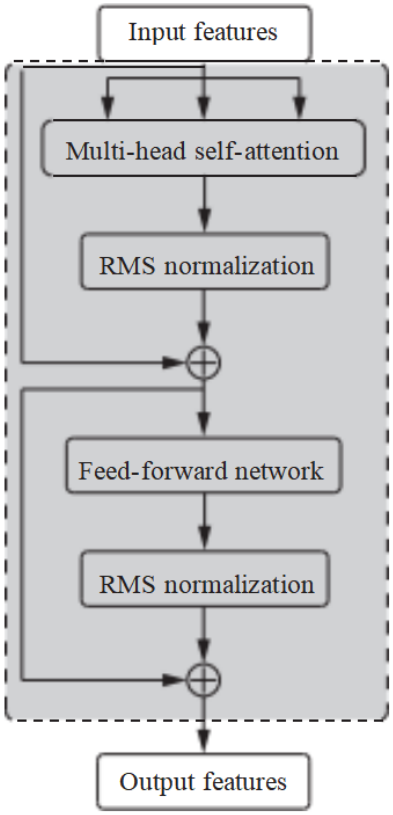

Figure 3: The structure of a single-layer transformer block

Each layer's multi-head self-attention mechanism first uses three linear layers to compute the Key, Query, and Value matrix features for each input. It then multiplies the Key and Query, scales the result, and applies the softmax function for normalization. Finally, it multiplies this by the Value matrix to generate each head's output, which is concatenated and fed through another linear layer to produce the final output of the multi-head self attention mechanism [20]:

$$\begin{cases} k_i, q_i, v_i = xw_{k_i}, xw_{q_i}, xw_{v_i} \\ z_i = \text{softmax}\left(\dfrac{q_i k_i^T}{\sqrt{(d_{k_i}}}\right) v_i \\ m_z = \text{Concat}\left(z_1, z_2, \cdots, z_h\right) w_m \end{cases} \tag{7}$$

where k_i , q_i, and v_i are the Key, Query, and Value matrix features of the ith head. x denotes the input feature. w_{ki} , w_{qi}, and w_{vi} are the linear layer weights for generating these features. z_i

is the attention feature of the *i*th head. dk_i is the first-dimension size of the Key matrix. m_z represents the output of the multi-head self-attention, and w_m is the output linear layer weight.

RMS normalization normalizes data features by calculating their root mean square. This scales data features to a unified range while maintaining their distribution and minimizing the impact of outliers. First, calculate the RMS of the input features. Then, normalize by dividing each input by its RMS. Finally, introduce two learnable parameters for scaling and shifting [21]:

$$r_i = \frac{x_i}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 + }} + \tag{8}$$

In the formula, xi and ri are the ith elements of the RMS normalized input and output features, respectively. n is the total number of elements, and is a small value. and are scaling and shifting parameters. Each layer's RMS normalization adds to the previous layer's input features to form a residual structure, preventing model degradation as depth increases.

The feed-forward network is a fully connected one with two linear transformations. It uses the ReLU activation function in between for nonlinear processing.

$$f = max(0; xw_1 + b_1)w_2 + b_2 \tag{9}$$

In the formula, x and f denote the input and output features of the feed-forward network, while w_1 and w_2 represent the weights, and b_1 and b_2 the biases of the two linear layers.

3.2 Dynamic knowledge self-distillation training strategy

Self-distillation is a distinct knowledge refinement technique. It enables models to self learn and enhance effectiveness. Unlike regular knowledge distillation that uses a pretrained complex teacher model, self-distillation eliminates this need. It simplifies training, cuts resource use, and acts as a strong regularizer. This improves the model's capacity for new data adaptation. Through self-learning, the model can extract its own knowledge and boost generalization without extra guidance [22].

The self-distillation framework in this paper consists ofa teacher backbone network and a student branch network. The teacher backbone is made up of multiple Transformer blocks and a teacher classifier, while the student branch has the first two Transformer blocks and a student classifier. During self-distillation, both networks are trained together. First, the outputs of both networks are transformed into probabilities via Softmax with a temperature coefficient to get soft labels and soft outputs. Then, KL divergence loss measures the difference between the teacher's soft labels and the student's soft outputs. The specific steps are as follows [23]:

$$\begin{cases} P_i = \mathrm{Softmax}\left(H_i^T, T\right) = \dfrac{\exp\left(H_i^T / T\right)}{\sum_j \exp\left(H_j^T\right) / T} \\ Q_i = \mathrm{Softmax}\left(H_i^S, T\right) = \dfrac{\exp\left(H_i^S / T\right)}{\sum_j \exp\left(H_j^S\right) / T} \\ L_{KL} = \sum_i P_i \log \dfrac{P_i}{Q_i} \end{cases} \tag{10}$$

where P^i and Q^i are the values of the ith elements in the probability distributions output by the teacher backbone and student branch networks, respectively. And the output features of the teacher and student networks. T is the temperature coefficient, and L_{KL} denotes the KL divergence loss.

During distillation, to enhance the student model's learning effectiveness, this paper constructs a temperature scaling module with two fully connected layers. This module dynamically adjusts the temperature parameter during distillation [24, 25]. It takes the outputs ofthe teacher backbone and student branch as input features to predict a temperature, which is then scaled based on the initial and maximum temperature to get the final temperature[26]:

$$T = Tini + Tup\,(Tpre) \tag{11}$$

where Tpre is the predicted temperature value from the temperature scaling module. Tini is the initial temperature, Tup is the upper limit of the highest temperature, and (.) is the nonlinear activation function.

$$\begin{cases} P_i = \mathrm{Softmax}\left(H_i^T\right) = \dfrac{\exp\left(H_i^T\right)}{\sum_j \exp\left(H_S^T\right)} \\ Q_i = \mathrm{Softmax}\left(H_i^S\right) = \dfrac{\exp\left(H_i^S\right)}{\sum_j \exp\left(H_j^S\right)} \\ L_{CE} = -\sum_i^C Y_i \log\left(P_i\right) - \sum_{i=1}^C Y_i \log\left(Q_i\right) \end{cases} \tag{12}$$

where LCE stands for cross entropy loss and C is the number of categories.

Finally, through a joint training process, the final loss function is designed as follows:

$$L = (1 -)L_{CE} + \quad L_{KL} \tag{13}$$

Through joint optimization, the student network simultaneously learns the soft labels of the teacher network and the implicit feature alignment of the contrastive branch, thereby improving detection performance. is a scaling factor. Gradually increasing during training progressivelyraises the difficulty of the student branch's learning task, promoting better knowledge acquisition from the teacher backbone. A binomial function is employed in this paper to control the rate of increase of [27, 28]:

$$(n) = s_{\text{start}} + \frac{n^2}{N^2}\left(end - \text{strat}\right) \tag{14}$$

4 Experimental Analysis

4.1 Data description and processing

This paper selects two distinct sample sets via a sensor system, each containing 5000 samples. For each dataset, 80% is allocated for training and 20% for testing. Each sample comprises six features: K1K2 door access signal (X_1), electronic lock drive signal (X_2), emergency stop voltage value (X_3), voltage access voltage value (X_4), harmonic distortion voltage value (X_5),

and harmonic distortion current value (X_6). As shown in Figure 4, the input feature dimension of each dataset is [1, 6], and the output dimension is (1), where [0] indicates a normal label and [1] a fault label. Before use, each dataset is normalized. This scales data to the 0-1 range for faster processing and converts dimensional parameters to non dimensional ones for effective interaction [29, 30]:

$$\mathcal{X} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{15}$$

The simulation experiment was performed on a computer equipped with the Core(TM)i9-14900KF processor and Nvidia GeForce RTX4090 GPU. The charging station fault detection model was established within the Python environment, leveraging Sklearn and Pytorch frameworks.

4.2 Feature reconstruction analysis

To assess how much the feature reconstruction module enhances the separability of original features, this paper employs t-SNE based feature-embedding visualization for dimensionality reduction. It compares the reconstruction effects under different window sizes m and determines the feature-combination strategy that can maximize class separability.

Firstly, permutation and combination are used to generate new features. The product of any m features among the original 6-dimensional limited features forms a new feature matrix. When m = 2, C_6^2 = 15 new features are generated; when m = 3, C_6^3 = 20 new features are generated; when m= 4, C_6^4 = 15 new features are generated. Then, tSNE is used for dimensional reduction to map the high dimensional features to a 2D space. The specific results are shown in Figure 5.

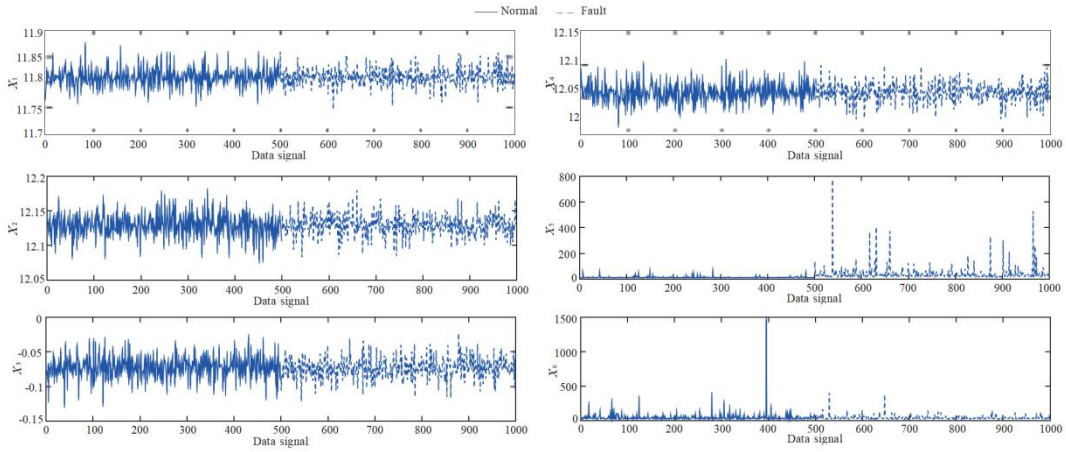

Figure 4: Charging pile fault detection sample parameters

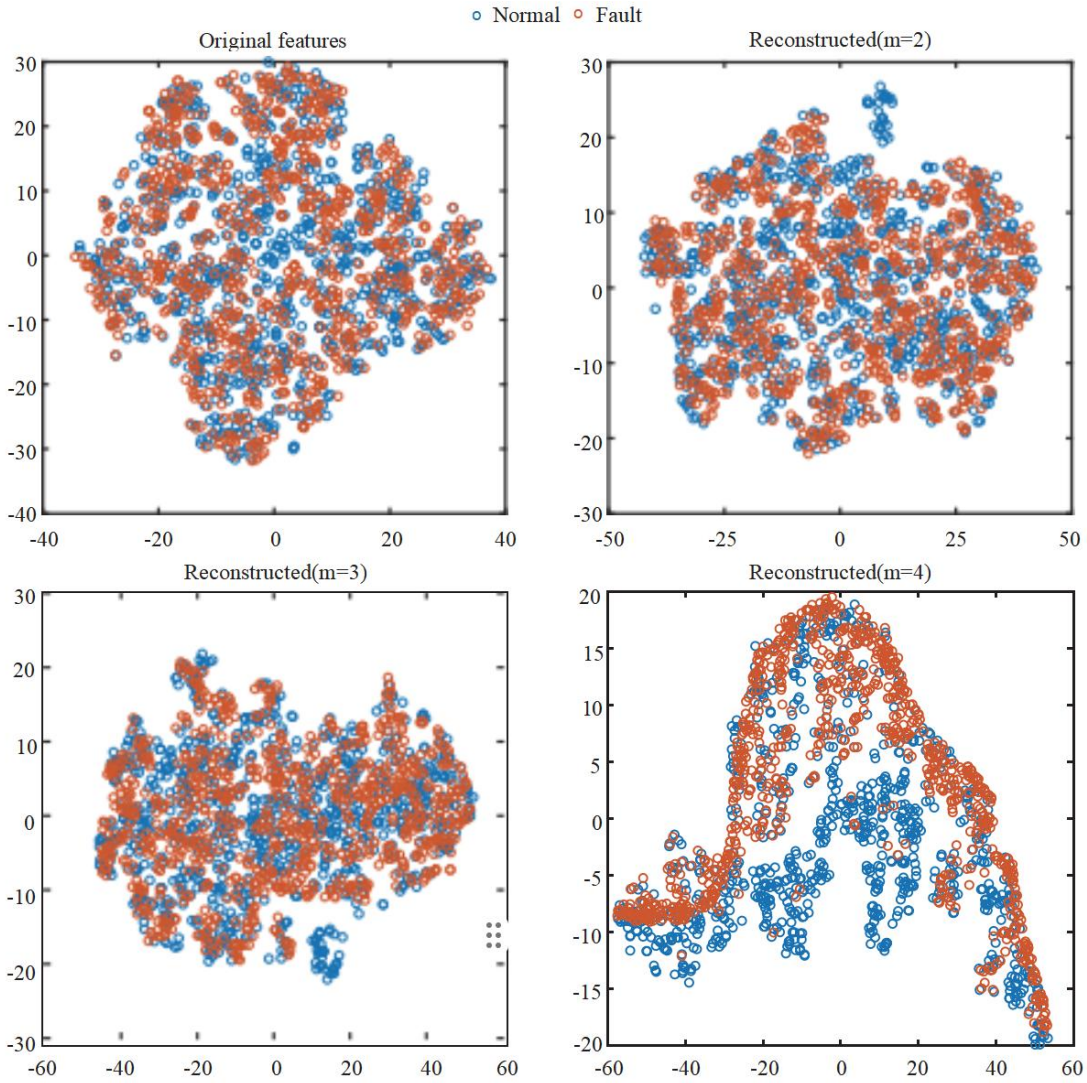

Figure 5: T-SNE verification feature reconstruction comparison visualization

5 Experimental Analysis

5.1 Data description and processing

As shown in Figure 5, when m = 4, the model can capture more complex feature interactions, and the class boundaries after reconstruction are the clearest. So, fourth order combination features are preferentially used to train the classification model.

5.2 Performance comparison experiment

This paper proposes a Dynamic Self-Distillation Transformer method for fault detection in charging stations with limited features. To verify its effectiveness, systematic comparative experiments are conducted with current mainstream methods:

1) XGBoost: Gradient boosting tree model. Parameters: max depth=6, number of estimators=200, learning rate=0.1.

2)1D-CNN: 1D convolutional neural network. Structure:

2 convolutional layers with kernel size=3, 1 pooling layer and 1 fully connected layer.

3) LSTM: Long Short-Term Memory network. Structure: 64-unit LSTM layer, 0.5 dropout and fully connected layer.

4) Transformer: Basic Transformer model. Structure: 4 encoder layers with 4-head attention and hidden size=128.

5) GAN+CNN: GAN-enhanced CNN. Generator is 3 fully connected layers and Discriminator is 1D-CNN.

All models use m = 4feature reconstruction, the same data split, and 5 fold cross validation on two real charging station datasets, with average results reported.

This paper assesses all detection models using recall, precision, and accuracy. Recall is the proportion of actual positive samples correctly identified by the model. A higher recall means fewer positive samples are missed. Precision is the proportion of true positives among all samples predicted as positive. When precision is high, a larger proportion of positive predictions are correct. Accuracy reflects the overall correctness of the model's predictions. Higher accuracy indicates that the model's predictions are closer to the ground truth:

$$\begin{cases} \text{Recall} = \dfrac{TP}{TP + FN} \\ \text{Precision} = \dfrac{TP}{TP + FP} \\ \text{Accuracy} = \dfrac{TP + TN}{TP + TN + FP + FN} \end{cases} \tag{16}$$

where *TP* is true positive, *FP* is false positive, *TN* is true negative and *FN* is false negative.

From the results in Table 1, XGBoost, as a tree based model, struggles to capture high order non-linear feature interactions. Its limited splitting depth in feature sparse scenarios leads to the poorest recognition performance. 1DCNN's convolutional kernels excel at local feature extraction but fail to model long term dependencies, giving it weaker global perception than sequential models but better recognition ability than XGBoost. LSTM and Transformer, with their strong global dependency modeling, outperform CNN significantly. The GAN+CNN data enhancement strategy backfires in feature scarce settings, generating samples that deviate from the true distribution and add extra noise.

The experimental results show that the model proposed in this paper has significant advantages in both accuracy and recall indicators, with an accuracy improvement of 3.2% and a recall improvement of 3.3%, fully verifying its high sensitivity and strong generalization ability to fault samples.

Table 1: Three evaluation metrics of different models on two datasets

Model	Dataset 1			Dataset 2		
	Recall(%)	Precision(%)	Accuracy(%)	Recall(%)	Precision(%)	Accuracy(%)
XGBoost	84.2±1.3	86.5±1.1	85.7±0.9	82.7±1.5	85.1±1.4	84.3±1.2
1D-CNN	87.6±0.8	89.3±0.7	88.5±0.6	86.1±1.1	88.7±0.9	87.6±0.8
LSTM	89.3±0.7	90.8±0.6	90.2±0.5	88.4±0.9	90.2±0.8	89.4±0.7
Transform	91.5±0.6	92.3±0.5	91.9±0.4	90.2±0.8	91.7±0.7	91.0±0.6
GAN+CNN	88.7±1.0	90.1±0.9	89.5±0.8	87.3±1.2	89.5±1.0	88.4±0.9
Our	94.2±0.4	95.1±0.3	94.7±0.3	93.5±0.5	94.8±0.4	94.2±0.4

This paper compares six models' computational efficiency for charging station fault detection across two datasets, assessing training time, per-sample testing time, and model size. As shown in Table 2, XGBoost has the smallest model size and fastest training due to its

simple tree based structure and no back propagation nature. LSTM's gated mechanism requires maintaining hidden states, increasing parameters and training time, and doesn't parallelize easily. Transformer's modeling of global dependencies involves computing associations between all positions, leading to high computational costs and the longest training time. The proposed method, using temperature scaling and progressive weight adjustment to distill knowledge from a deep Transformer into a lightweight student network, improves computational efficiency by 16.7% over the second best 1D-CNN and reduces model size by 56% compared to LSTM, offering a high precision, low latency solution for charging station fault detection.

Table 2: The computational efficiency of different models

Model	Dataset 1			Dataset 2		
	Training time(s)	Testing time(s)	Model size(MB)	Training time(s)	Testing time(s)	Model size(MB)
XGBoost	3.2	0.002	0.8	3.5	0.002	0.9
1D-CNN	18.7	0.006	2.1	19.3	0.006	2.2
LSTM	25.4	0.009	3.8	26.8	0.009	3.9
Transform	42.6	0.013	8.7	43.9	0.014	8.7
GAN+CNN	112.5	0.011	10.3	118.3	0.012	10.3
Our	38.2	0.005	1.7	39.6	0.005	1.7

5.3 Ablation experiment

The model in this article consists of two core modules: feature reconstruction module and dynamic knowledge distillation module. The contribution of each module to the model performance is verified through ablation experiments. The specific settings of each ablation model are shown in Table 3.

Figure 6 compares the true values and recognition results of the complete model and each ablation model under two datasets. According to the results shown in Figure 6, the recognition accuracy of ablation model 1 with missing feature reconstruction module significantly decreased, indicating that this module can improve model performance by mining effective feature information. Removing the dynamic knowledge distillation module significantly reduces the accuracy of ablation model 2, and removing the shallow student network also significantly reduces the accuracy of ablation model 3, indicating that deep transformers have strong nonlinear fitting capabilities, and knowledge distillation can assist shallow networks in learning key knowledge of deep models. The ablation model 4 without a dynamic adjustment mechanism performs worse than the complete model, verifying the effectiveness of dynamic knowledge distillation in improving detection accuracy.

Table 3: Detailed description of the ablation models

Comparison model	Detail description
Ablation Model 1	Remove the feature reconstruction module on the basis of the original model
Ablation Model 2	Only the teacher backbone network is used for feature extraction
Ablation Model 3	Only the student branch is used for feature extraction
Ablation Model 4	Remove the temperature module on the basis of the original model

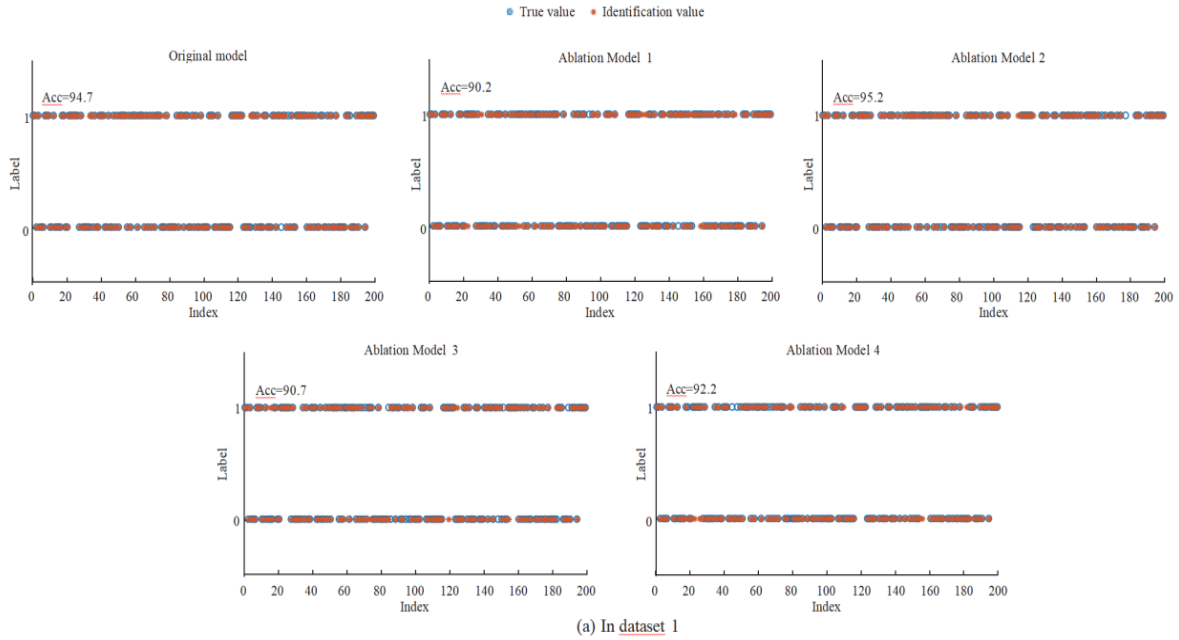

Figure 6: The original model was tested and visualized with different ablation models

5.4 Verification experiment of dynamic distillation

To verify how the adaptive mechanism of temperature coefficient in dynamic self-distillation boosts model performance in charging station fault detection, the experiment compares dynamic and fixed temperature scaling strategies. It trains for 100 epochs on two charging station datasets, tracking τ changes and their effects on training and loss validation accuracy. The results are shown in Figure 8.

As shown in Figure 7, dynamic τ drops from 1.92 to 0.68 in three stages. The initial high temperature softens the teacher's output distribution, easing knowledge transfer. The mid training moderate temperature focuses on key features.

The final low temperature aligns with the true label distribution, suppressing overfitting noise. The dynamic strategy hits 92.2% accuracy by epoch 40, a 20% faster convergence than the fixed one, and reaches a higher final accuracy of 94.7%.

This also compared fixed λ with values of 0.2, 0.4, 0.6, and 0.8 to a binomially increasing λ to show dynamic knowledge distillation's advantages. Figure 8 presents the results.

It is evident that the model achieves the highest detection accuracy when using binomially increased λ. This indicates that incrementally increasing λ allows the student model to more effectively learn from the teacher model. Compared to fixed λ, this approach significantly reduces model error.

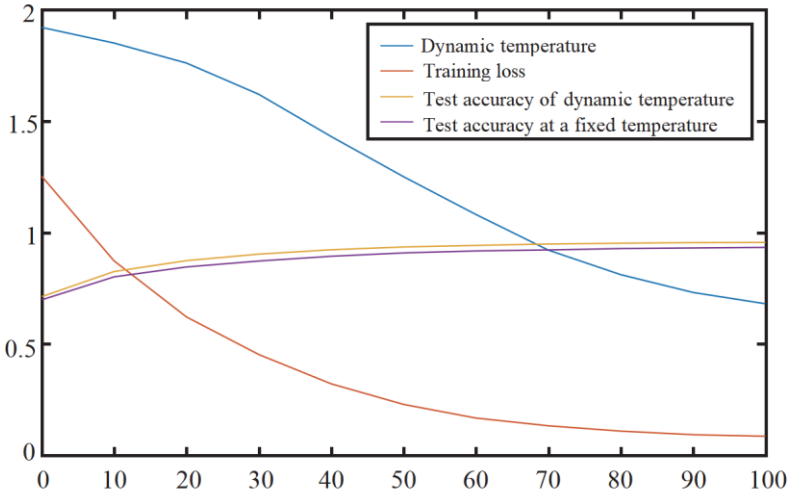

Figure 7: The variation of performance results under different temperature strategies

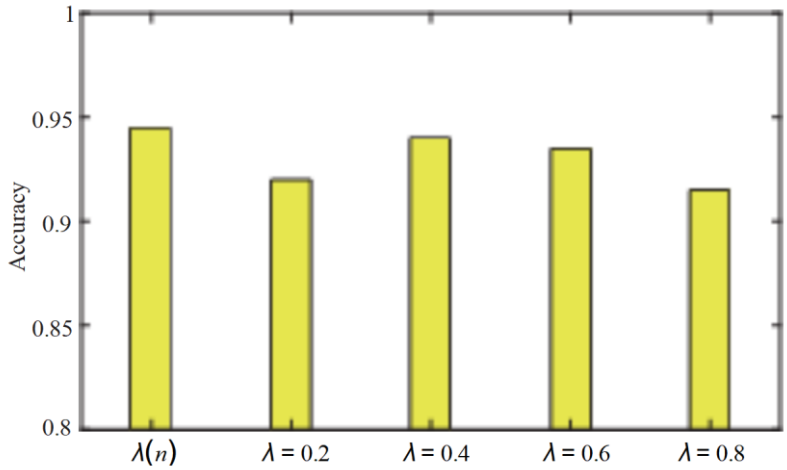

Figure 8: Accuracy as λ changes

6 Conclusion

To address the challenge of limited feature charging station fault detection, this study proposes a dynamic selfdistilled Transformer based solution. It first rearranges and combines limited input features to enhance their usefulness. Then, a dynamic knowledge self-distillation strategy trains the Transformer network for in depth feature mining. This approach effectively compresses knowledge from deeper networks into shallower ones, ensuring high detection accuracy and low model complexity. Experiments on two real world datasets show that the model outperforms current state of the art models in recognition accuracy. It also confirms that knowledge distillation helps shallow models learn from deeper ones, benefiting subsequent model deployment.

References

[1] Shufian A, Hannan N, Emon M M H, et al. Automatic Fault Detection and Analysis of Electric Vehicle Charging Station by Machine Learning[C]//2024 IEEE Region 10

Symposium (TENSYMP). IEEE, 2024: 1-6.

[2] Alshareef S M. Voltage sag assessment, detection, and classification in distribution systems embedded with fast charging stations[J]. IEEE Access, 2023, 11: 89864-89880.

[3] Gupta A, Sarangi S, Singh A K. Wavelet based enhanced fault detection scheme for a distribution system embedded with electric vehicle charging station[C]//2023 5th International Conference on Power, Control & Embedded Systems (ICPCES). IEEE, 2023: 1-6.

[4] Hosseini S A, Taheri B, Sadeghi S H H, et al. A deep learning model for fault detection in distribution networks with high penetration of electric vehicle chargers[J]. e-Prime-Advances in Electrical Engineering, Electronics and Energy, 2024, 10: 100845.

[5] Zurek-Mortka M, Szymanski J R. The resistive ground fault of PWM voltage inverter in the EV charging station[J]. Scientific reports, 2021, 11(1): 21236.

[6] Deb N, Singh R, Brooks R R, et al. A review of extremely fast charging stations for electric vehicles[J]. Energies, 2021, 14(22): 7566.

[7] Salehimehr S, Miraftabzadeh S M, Brenna M. A novel machine learning-based approach for fault detection and location in low-voltage dc microgrids[J]. Sustainability, 2024, 16(7): 2821.

[8] Faraji H, Khorsandi A, Hosseinian S H. Multi-level coordinated control of islanded DC microgrid integrated with electric vehicle charging stations with fault ride-through capability[J]. Journal of Cleaner Production, 2023, 420: 138372.

[9] Hussain A, Yadav A, Ravikumar G. Anomaly detection using bi-directional long short-term memory networks for cyber-physical electric vehicle charging stations[J]. IEEE Transactions on Industrial Cyber-Physical Systems, 2024.

[10] Bayrak G, Yılmaz A, Çakmak R. A new Fuzzy&Wavelet-based adaptive thresholding method for detecting PQDs in a hydrogen and solar-energy powered EV charging station[J]. International Journal of Hydrogen Energy, 2023, 48(18): 6855-6870.

[11] Khalid M, Ahmad F, Panigrahi B K. Design, simulation and analysis of a fast charging station for electric vehicles[J]. Energy Storage, 2021, 3(6): e263.

[12] Joga S R K, SaiPrakash C. A novel method to detect and classify High Impedance Fault in EV integrated distribution system[C]//2023 IEEE International Conference on Power Electronics, Smart Grid, and Renewable Energy (PESGRE). IEEE, 2023: 1-6.

[13] Kosuru V S R, Kavasseri Venkitaraman A. A smart battery management system for electric vehicles using deep learning-based sensor fault detection[J]. World Electric Vehicle Journal, 2023, 14(4): 101.

[14] Abu-Nassar A M, Morsi W G. Early detection of cyber-physical attacks on electric vehicles fast charging stations using wavelets and deep learning[J]. IEEE Transactions on Industrial Cyber-Physical Systems, 2024, 2: 220-231.

[15] Nazih Y, Abdel-Moneim M G, Aboushady A A, et al. A ring-connected dual active bridge based DC-DC multiport converter for EV fast-charging stations[J]. IEEE Access, 2022, 10: 52052-52066.

[16] Olcay K, Çetinkaya N. Analysis of the electric vehicle charging stations effects on the electricity network with artificial neural network[J]. Energies, 2023, 16(3): 1282.

[17] Rajendran G, Vaithilingam C A, Naidu K, et al. Open switch fault-tolerant VOC-PI controller based Vienna rectifier for EV charging stations[C]//2021 International Conference on Electrical Engineering and Informatics (ICEEI). IEEE, 2021: 1-5.

[18] Herbst D, Fürnschuß M, Reichel P, et al. Challenges and related solutions for periodic verification of DC electric vehicle charging stations[C]//CIRED Porto Workshop 2022: E-mobility and power distribution systems. IET, 2022, 2022: 113-117.

[19] Choudhary A, Fatima S, Panigrahi B K. State-of-the-art technologies in fault diagnosis of electric vehicles: A component-based review[J]. IEEE Transactions on Transportation Electrification, 2022, 9(2): 2324-2347.

[20] Balasundar C, Sundarabalan C K, Srinath N S, et al. Effect of fault ride through capability on electric vehicle charging station under critical voltage conditions[J]. IEEE transactions on transportation electrification, 2022, 8(2): 2469-2478.

[21] Franzese P, Patel D D, Mohamed A A S, et al. Fast DC charging infrastructures for electric vehicles: Overview of technologies, standards, and challenges[J]. IEEE Transactions on Transportation Electrification, 2023, 9(3): 3780-3800.

[22] Kivelä T, Abdelawwad M, Sperling M, et al. Functional Safety and Electric Vehicle Charging: Requirements Analysis and Design for a Safe Charging Infrastructure System[C]//VEHITS. 2021: 317-324.

[23] Aljohani T, Almutairi A. Modeling time-varying wide-scale distributed denial of service attacks on electric vehicle charging Stations[J]. Ain Shams Engineering Journal, 2024, 15(7): 102860.

[24] Watil A, Chojaa H. Enhancing grid-connected PV-EV charging station performance through a real-time dynamic power management using model predictive control[J]. Results in Engineering, 2024, 24: 103192.

[25] Aljohani T, Almutairi A. Modeling time-varying wide-scale distributed denial of service attacks on electric vehicle charging Stations[J]. Ain Shams Engineering Journal, 2024, 15(7): 102860.

[26] Watil A, Chojaa H. Enhancing grid-connected PV-EV charging station performance through a real-time dynamic power management using model predictive control[J]. Results in Engineering, 2024, 24: 103192.

[27] Ejenakevwe K A, Song L. Investigation of smart thermostat fault detection and diagnosis potential for air-conditioning systems using a Modelica/EnergyPlus co-simulation approach[J]. Energy and Buildings, 2024, 309: 114053.

[28] Grcić I, Pandžić H. Artificial neural network for high-impedance-fault detection in DC microgrids[C]//2023 IEEE PES Conference on Innovative Smart Grid Technologies-Middle East (ISGT Middle East). IEEE, 2023: 1-5.

[29] Adetunji K E, Hofsajer I W, Abu-Mahfouz A M, et al. A novel dynamic planning mechanism for allocating electric vehicle charging stations considering distributed generation and electronic units[J]. Energy Reports, 2022, 8: 14658-14672.

[30] Altaf M, Yousif M, Ijaz H, et al. PSO-based optimal placement of electric vehicle charging stations in a distribution network in smart grid environment incorporating backward forward sweep method[J]. IET Renewable Power Generation, 2024, 18(15): 3173-3187.